Intelligent Information Systems Institute
AFOSR Grant FA9550-04-1-0151

Final Performance Report

Period Covered March 2004 through February 2008
June, 2008

Carla P. Gomes
Computing and Information Science
Cornell University
Ithaca, NY

(607) 255-9189 (phone)
(607) 255-4428 (fax)
gomes@cs.cornell.edu

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 27, 2008	Final Report 3/1/04-2/28/08

4. TITLE AND SUBTITLE
Intelligent Information Systems Institute

5. FUNDING NUMBERS
FA9550-04-1-0151

6. AUTHOR(S)
Gomes, Carla

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Cornell University
Ithaca, NY 14853

8. PERFORMING ORGANIZATION REPORT NUMBER
45158

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NL
4015 Wilson Boulevard, Room 713
Arlington, VA 22203-1954

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-SR-AR-TR-08-0383

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; distribution is Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The Intelligent Information Systems Institute began operation in December of 2000. Its mandate is threefold: To perform and stimulate research in compute- and data-intensive methods for intelligent decision making systems; to foster collaborations between Cornell researchers, AFRL, in particular IF, and the scientific community; and to play a leadership role in the research and dissemination of the core areas of the institute. IISI supports basic research within the Faculty of Computing and Information Science (FCIS). The institute defies the traditional barriers between disciplines, constituting a catalyst to new collaborative, interdisciplinary projects, with a focus on computational intensive and data intensive methods arising in large-scale decision systems. IISI promotes a cross-fertilization of approaches from different disciplines, including computer science, engineering, operations research, economics, mathematics, statistics, and physics. During the period covered by this grant the institute focused on three themes: 1) Computational Complexity, Typical Case Analysis, Problem Structure, and Connections to Statistical Physics, 2) Mathematical and Computational Foundations of Networks, and 3) Autonomous and Distributed Agents, generating over 100 research publications, including invited perspectives in Science and Nature that disseminate to a broader science audience various research results such as on phase transitions, combinatorial optimization, heavy-tailed behavior in computation, and constraint reasoning.

14. SUBJECT TERMS

15. NUMBER OF PAGES
19

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

1. Executive Summary

The Intelligent Information Systems Institute began operation in December of 2000. Its mandate is threefold: To perform and stimulate research in compute- and data-intensive methods for intelligent decision making systems; to foster collaborations between Cornell researchers, AFRL, in particular IF, and the scientific community; and to play a leadership role in the research and dissemination of the core areas of the institute. IISI supports basic research within the Faculty of Computing and Information Science (FCIS). The institute defies the traditional barriers between disciplines, constituting a catalyst to new collaborative, interdisciplinary projects, with a focus on computational intensive and data intensive methods arising in large-scale decision systems. IISI promotes a cross-fertilization of approaches from different disciplines, including computer science, engineering, operations research, economics, mathematics, statistics, and physics. During the period covered by this grant the institute focused on three themes: 1) Computational Complexity, Typical Case Analysis, Problem Structure, and Connections to Statistical Physics, 2) Mathematical and Computational Foundations of Networks, and 3) Autonomous and Distributed Agents, generating over 100 research publications, including invited perspectives in Science and Nature that disseminate to a broader science audience various research results such as on phase transitions, combinatorial optimization, heavy-tailed behavior in computation, and constraint reasoning.

2. Research Progress

We describe our research accomplishments and findings, since our last performance report in September 2006, considering the two major research themes:

Theme 1 - Controlling Computational Cost
Theme 2 - Mathematical and Computational Foundations of Complex Networks

Theme 1 - Controlling Computational Cost

This theme encompasses foundational work on "typical" computational complexity, study of structure, and randomization in hard computational problems. This research combines formal analysis and design of optimization techniques with the study of a range of applications to large-scale computational problems, such as distributed networks, autonomous agents, and combinatorial auctions, applications such as planning and scheduling, autonomous distributed agents, and combinatorial auctions. Key research topics of our work are (1) the integration of optimization concepts from artificial intelligence, and operations research, in particular mathematical programming and constraint programming, (2) foundational work on "typical" computational complexity, (3) the study of the impact of structure on problem hardness of hard computational problems, (4) the use of randomization techniques to improve the performance of search

methods, and (5) global optimization through machine learning and response surface methods. Below some findings are highlighted.

Tradeoffs in the Complexity of Backdoor Detection

Capturing and exploiting problem structure is a key to solving large real-world combinatorial problems. For example, several interesting tractable classes of combinatorial problems have been identified by restricting the constraint language used to characterize such problem instances. Well-known cases include 2CNF, Horn, Linear Programming (LP), and Minimum Cost Flow problems (MCF). In general, however, such restricted languages are not rich enough to characterize complex combinatorial problems. A very fruitful and prolific line of research that has been pursued in the study of combinatorial problems is the identification of various structural properties of instances that lead to efficient algorithms. Ideally, one prefers structural properties that are "easily" identifiable, such as from the topology of the underlying constraint graph. As an example, the degree of acyclicity of a constraint graph, measured using various graph width parameters, plays an important role with respect to the identification of tractable instances — it is known that an instance is solvable in polynomial time if the treewidth of its constraint graph is bounded by a constant. Interestingly, even though the notion of bounded treewidth is defined with respect to tree decompositions, it is also possible to design algorithms for constraint satisfaction problems of bounded (generalized) hypertree width that do not perform any form of tree decomposition. Other useful structural properties consider the nature of the constraints, such as their so-called functionality, monotonicity, and row convexity. Another approach for studying combinatorial problems developed by our research group focuses on the role of hidden structure as a way of analyzing and understanding the efficient performance of state-of-the-art constraint solvers on many real-world problem instances. One example of such hidden structure is a backdoor set, i.e., a set of variables such that once they are instantiated, the remaining problem simplifies to a tractable class. Note that the notion of tractability in the definition of backdoor sets is not necessarily syntactically defined: it may often be defined only by means of a polynomial time algorithm, such as unit propagation. In fact, the notion of backdoor sets came about as a way of explaining the high variance in performance of state-of-the-art SAT solvers, in particular heavy-tailed behavior, and as a tool for analyzing and understanding the efficient performance of these solvers on many real-world instances, in which the propagation mechanisms of fast "sub-solvers" play a key role. In this work the emphasis was not so much on efficiently identifying backdoor sets, but rather on the fact that many real-world instances have surprisingly small sets of backdoor variables and that once a SAT solver instantiates these variables, the rest of the problem is solved easily. In this context, randomization and restarts play an important role in searching for backdoor sets.

Even though variable selection heuristics, randomization, and learning in current SAT/CSP solvers are quite effective at finding relatively small backdoors in practice, finding a smallest backdoor is in general intractable in the worst case. This intractability result assumes that the size of the smallest backdoor is unknown and can grow arbitrarily

with n. However, if the size of the backdoor is small and fixed to k, one can search for the backdoor by considering all $C_{n,k}$ subsets of k variables and all 2^k truth assignments to these candidate variables. This is technically a polynomial time process for fixed k, although for moderate values of k the run time becomes infeasible in practice. Can one do better? This is a question considered in the area of fixed-parameter complexity theory. A problem with input size n and a parameter k is called fixed-parameter tractable w.r.t. k if it can be solved in time $O(f(k)n^c)$ where f is any computable function and c is a constant. Note that c does not depend on k, meaning that one can in principle search fairly efficiently for potentially large backdoors if backdoor detection for some class is shown to be fixed parameter tractable. Indeed, Nishimura, Ragde, and Szeider showed that detecting strong backdoors w.r.t. the classes 2CNF and Horn is NP-complete but, interestingly, fixed-parameter tractable. This result for 2CNF and Horn formulas exploits the equivalence between (standard) strong backdoors and "deletion" backdoors, i.e., a set of variables that once deleted from a given formula (without simplification) make the remaining formula tractable. Note, however, that this result is only w.r.t. the tractable classes of pure 2CNF/Horn.

In particular, certain kinds of obvious inconsistencies are not detected in these classes, such as having an empty clause in an arbitrary formula — clearly, any basic solver detects such inconsistencies. We show that such a seemingly small feature increases the worst-case complexity of backdoor identification, but, perhaps more importantly, can dramatically reduce the size of the backdoor sets. More specifically, we have proved that strong Horn- and 2CNF-backdoor identification becomes both NP- and coNP-hard, and therefore strictly harder than NP assuming NP ≠ coNP, as soon as empty clause detection is added to these classes. This increase in formal complexity has however also a clear positive aspect in that adding empty clause detection often considerably reduces the backdoor size. For example, in certain graph coloring instances with planted cliques of size 4, while strong Horn-backdoors involve ≈ 67% of the variables, the fraction of variables in the smallest strong backdoors w.r.t. mere empty clause detection converges to 0 as the size of the graph grows. Encouraged by the positive effect of slightly extending our notion of Horn-backdoor, we also consider backdoors w.r.t. RHorn (renamable Horn), UP (unit propagation), PL (pure literal rule), UP+PL, and SATZ. For each of these notions, we have shown on a variety of domains that the corresponding backdoors are significantly smaller than pure, strong Horn-backdoors. For example, we considered the smallest deletion RHorn-backdoors. We developed a 0-1 integer programming formulation for finding such optimal backdoors, and showed experimentally that they are in general smaller than strong Horn-backdoors. In particular, in the graph coloring domain, while strong Horn-backdoors correspond to ≈ 67% of the variables, deletion RHorn-backdoors correspond to only ≈ 17% of the variables. More interestingly, when considering real-world instances of a car configuration problem, while strong Horn-backdoor sets vary in size between 10-25% of the variables, deletion RHorn-backdoor sets vary only between 3-8%. At a higher level, our results show that the size of backdoors can vary dramatically depending on the effectiveness of the underlying simplification and propagation mechanism. For example, as mentioned earlier, empty clause detection can have a major impact on backdoor size. Similarly, Horn versus RHorn has an impact. We also showed that there can be a substantial

difference between deletion backdoors, where one simply removes variables from the formula, versus strong backdoors, where one factors in the variable settings and considers the propagation effect of these settings. Specifically, we contrast deletion RHorn-backdoors with strong RHorn-backdoors. We proved by construction that there are formulas for which deletion RHorn-backdoors are exponentially larger than the smallest strong RHorn-backdoors. Finally, despite the worst-case complexity results for strong backdoor detection, we showed that Satz-Rand is remarkably good at finding small strong backdoors on a range of experimental domains. For example, in the case of our graph coloring instances, the fraction of variables in a small strong SATZ-backdoor converges to zero as the size of the graph grows. For the car configuration problem, strong SATZ-backdoor sets involve 0-0.7% of the variables. We also considered synthetic logistics planning instances over n variables that are known to have strong UP-backdoors of size log n. For all these instances, the size of the strong SATZ-backdoor sets is either zero or one. In contrast, the size of deletion RHorn-backdoors corresponds to over 48% of the variables, increasing with n. We also considered instances from game theory for which one is interested in determining whether there is a pure Nash equilibrium. For these instances, while strong Horn-backdoors and deletion RHorn-backdoors involve ≈ 68% and ≈ 67% of the variables, respectively, strong SATZ-backdoors are surprising small at less than 0.05% of the variables. These results show that real-world SAT solvers such as Satz are indeed remarkably good at finding small backdoors sets. At a broader level, this work suggests that the study of structural notions that lead to efficient algorithms for combinatorial problems should consider not only "easily" identifiable properties, such as being Horn, but also properties that capture key aspects of state-of-the-art constraint solvers, such as unit propagation and pure literal rule.

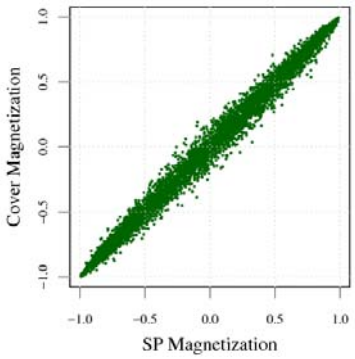
Understanding, Improving, and Exploiting Propagation Methods for SAT

In this project, Kroc, Sabharwal, and Selman investigate an exciting new technique called Survey Propagation (SP) for solving hard random combinatorial problems, with the goal of (a) providing insights into its somewhat mysterious and dramatic success, (b) using this study to further improve the related Belief Propagation (BP) family of algorithms, and (c) exploiting these improved algorithms for harder combinatorial tasks such as solution counting. This work was done in the context of Boolean constraint reasoning and probabilistic reasoning, which are inter-related fields and have numerous applications.

Survey Propagation was discovered by Mezard, Parisi, and Zecchina in 2002, and is so far the only known method successful at solving random Boolean satisfiability (SAT) problems with 1 million variables and beyond in near-linear time in the hardest region. Unlike the usual backtrack-based (DPLL) and local search based methods for solving SAT, the SP method is quite radical in that it tries to approximate certain marginal probabilities related to the set of satisfying assignments. It then iteratively assigns values to variables with the most extreme probabilities, thereby "decimating" the formula and making it simpler. In effect, the algorithm behaves like the usual backtrack search

methods for SAT (DPLL-based), which also assign variable values incrementally in an attempt to find a satisfying assignment. However, quite surprisingly, SP almost never has to backtrack. In other words, the "heuristic guidance" from SP is almost always correct and solves the problem in one shot. Note that, interestingly, computing marginals on satisfying assignments is actually believed to be much harder than finding a single satisfying assignment (#P-complete vs. NP-complete). Nonetheless, SP is able to efficiently approximate certain marginals and uses this information to successfully find a satisfying assignment.

This dramatic success of SP on the largest random problems modern SAT solvers can handle raises some natural questions: What is the combinatorial object that the iterative, mutually recursive equations of SP compute? What is the key aspect that makes SP different from the dozens of heuristics and techniques that have been tried by SAT researchers in the last 15-20 years? Can we de-mystify SP and present it purely as a general technique, say BP, applied to a well-defined combinatorial constraint satisfaction problem? Can the information computed by SP be of any use in structured problems rather than random SAT instances?



The recent work of Kroc, Sabharwal, and Selman answers these questions to various extents. It was discovered by Braunstein and Zecchina (later extended by Maneva, Mossel, and Wainwright) that SP equations are equivalent to BP equations for obtaining marginals over a special class of combinatorial objects called covers. Intuitively, a cover provides a representative generalization of a cluster of satisfying assignments. However, subsequent experiments suggested that covers do not exist, thus dismissing the promising line of thinking involving covers. Using an extensive empirical study, Kroc *et al.* gave strong evidence that not only do covers exist in significant numbers, SP is remarkably good at computing marginals over these covers; the figure depicts the alignment of the marginals computed by SP against (sampled) cover marginals. They further demonstrated that on random satisfiability problems, cover marginals are very close to solution marginals, differing only in the extreme marginals regime and in such a way that covers can be seen as more "conservative" than solutions. This work provides the first clear demonstration of what exactly the mysterious SP equations compute on large problem instances.

They also discovered that the variables set by SP follow a distinctive pattern not seen in any other well-known SAT solver heuristic, namely, avoiding constraint propagation (also known as unit propagation). While most know heuristics deliberately try to set variables that cause constraint propagation because all but one variable in a constraint is set the 'wrong' way, SP does exactly the opposite. This behavior can be related to SP's preference for certain "pure" variables such that setting them not only keeps the problem satisfiable but also makes it significantly easier to solve. Understanding the role of these special variables appears to be the key to explaining the success of SP.

In addition to the empirical results, Kroc *et al.* also revisited the derivation of the SP equations themselves, with the goal of presenting the derivation in an insightful form purely within the realm of combinatorial constraint satisfaction problems (CSPs). They described how one can reformulate in a natural step-by-step manner the problem of finding a satisfying assignment into one of finding a cover, by considering related factor graphs on larger state spaces. The BP equations for this reformulated problem are exactly the SP equations for the original problem.

Finally, the insights obtained during this process led Kroc *et al.* to propose a modified form of iterative BP equations that address a critical bottleneck when trying to apply these techniques to non-random problems, namely, the lack of convergence of the iterative process. The modified parameterized variant converges more easily as the parameter is reduced, with guaranteed convergence to a local computation when the parameter becomes zero. With this 'convergent BP,' they demonstrated how the sampling-based solution counting method introduced earlier by researchers at IISI can be extended to use guidance from BP in order to obtain improved results with orders of magnitude speed-up in practice. This is one of the few successful examples of applying propagation-based techniques to efficiently solve hard combinatorial problems substantially better than alternative existing methods.

Counting and Sampling in Combinatorial Solution Spaces

In this ongoing project, Gomes, van Hoeve, Hoffmann, Sabharwal, and Selman consider various aspects of the two related problems of counting and uniformly sampling from the solutions of hard combinatorial problems modeled as propositional formulas or general constraint satisfaction problems (CSPs). Many discrete real-world problems, such as those from hardware and software verification, planning, scheduling, algebra, and design automation, can be naturally and efficiently modeled as CSPs, and often as propositional formulas: True/False assignments to the variables of such a formula satisfy the formula if and only if that assignment corresponds to a solution of the original problem being modeled. While propositional satisfiability algorithms or SAT solvers researched extensively in the last 15 years attempt to find *one* such satisfying assignment (which is already a challenging NP-complete problem), the goal of this project is to develop techniques that provide much more information than a single solution, namely, the number of solutions or a single solution sampled uniformly from the space of all solutions. Such information is known to be extremely useful in probabilistic inference problems, such as Bayesian networks reasoning and Markov logic networks, and in combinatorial designs.

The obvious difficulty of these tasks stems from the fact that a constraint satisfaction problem with n variables can have anywhere from 0 to 2^n satisfying assignments, even when the variables are binary, that is, take only two values. Testing each potential assignment is clearly infeasible even for moderate values of n. While state-of-the-art constraint solvers can often find one satisfying assignment relatively quickly, the

heuristics employed by them are not designed to be efficient at exploring every part of the search space or sampling uniformly from the set of solutions. The challenge addressed in this project is to utilize the techniques used by modern SAT and CP solvers for the harder questions of solution counting and uniform sampling.

The key new contribution of Gomes *et al.* is an extension of their recent novel approach from last year that uses problem reformulation using the so-called random XOR or parity constraints for both solution counting and sampling. They provide (probabilistic) guaranteed lower and upper bounds in the case of counting and guaranteed near-uniform samples in the case of sampling. The surprising and distinguishing strength of their strategy is that they are able to use *any* state-of-the-art complete constraint solver off the shelf without any modifications whatsoever for both counting and sampling – problems computationally much more difficult than accomplishing a constraint solver's goal of finding a single solution.

The central idea of their main approach is to repeatedly add randomly chosen XOR or parity constraints on the problem variables to the input formula and feed the result to a constraint solver. At a very high level, each random XOR constraint is a streamlining constraint that cuts the search space approximately in half (assuming the domains are binary), irrespective of the distribution of solutions in the search space. Furthermore, if the solutions space is cut purely randomly in halves each time and one is left with a single solution in the end, this surviving solution must be a uniform sample.

Gomes *et al.* extensively evaluated the effect of the size or length of the XOR constraints on the quality of the solution count obtained. The key finding here was that although theoretically one needs 'large' XOR constraints for the technique to work at its best, in practice much 'smaller' XORs suffice to produce good results. Such small XORs are much easier to implement than large XORs, and are therefore highly preferred in practical implementations of the technique.

They extended this counting framework to general constraint satisfaction problems (CSPs) beyond disjunctive constraints over Boolean variables found in SAT. They created and evaluated two strategies for extending the XOR-based counting method: (a) create binary shadow variables to which regular binary XOR constraints are added, and (b) use generalized XOR constraints defined as the sum modulo the largest domain size being restricted to a randomly chosen right-hand-side. In both cases, the rich structure of the constraint programming (CP) framework allows one to explore more complex propagation and filtering techniques than what the SAT framework supports, such as Gaussian elimination based complete filtering and watched-literals based individual filtering hybridized with brute-force expansion. The experiments revealed that this generalized counting method works very well on problems with a rich structure that is simply not suitable for translating into the Boolean satisfiability framework. The approach was also shown to be competitive against integer programming (IP) based counting methods.

Multi-Agent Scheduling

van Hoeve and Lombardi studied multi-agent scheduling: agents need to be assigned tasks to execute, such that a collaborative goal is achieved with maximum quality. Important aspects of the particular problems under consideration are the nonlinear objective function (the goal), and the dynamic duration of activities. The latter stems from interaction between agents: actions performed by one agent may alleviate or aggravate the tasks performed by another agent. These aspects magnify the already present complexity of multi-agent scheduling problems. The approach taken by van Hoeve is to model and solve these problems within a constraint programming environment. He showed that constraint programming provides a natural modeling language to express the many complex relations. Moreover, the underlying solving technology of constraint programming, including domain filtering algorithms, constraint propagation, and specialized search procedures, allowed solving many problem instances to optimality. These results are reported in a paper to be presented at the Nineteenth Conference on Innovative Applications of Artificial Intelligence.

An important feature of the abovementioned system for solving multi-agent problems is the ability to decompose a given problem into independently solvable sub-problems. Together with M. Lombardi (University of Bologna, Italy and visitor of IISI in the past year), van Hoeve developed a theory to safely apply such decomposition in the context of general constraint optimization problems. This theory was also applied to the multi-agent scheduling problems, in which case speed-ups of several orders of magnitude were obtained. A paper describing this work has been submitted and is currently under review.

Connections in Networks

J. Conrad, C.P. Gomes, W.-J. van Hoeve, A. Sabharwal, and J. Suter studied the problem of identifying "connections in networks". One important application of this research is the ability to create wildlife corridors to enable species to travel from one reserve area to another. Other applications can be found, e.g., in social networks. In such problems, the goal is optimize the utility (based on connectedness), while keeping the cost below a given budget. The approach taken by Conrad et al. was to analyze the trade-off between utility/connectedness and the cost. More abstractly, the problem consists of two competing aspects: feasibility versus optimization. Conrad et al. discovered interesting phase transitions (from feasible to infeasible) and corresponding computational hardness peaks for these types of problems. Interestingly, proving infeasibility was shown to be much harder than proving optimality in a computationally hard region of the problem space. These results were published in a paper presented at the Fourth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems.

Filtering

van Hoeve, G. Pesant , L.-M. Rousseau (both University of Montreal) and A. Sabharwal developed several domain filtering algorithms for the sequence constraint. This constraint can be applied, e.g., to model and solve nurse-rostering problems, and to design production lines for car-manufacturing. Previous filtering methods for this constraint did not guarantee to establish complete filtering of the domains. In a paper presented at the Twelfth International Conference on Principles and Practice of Constraint Programming, van Hoeve et al. presented three new filtering algorithms, including the first algorithm achieving complete filtering in polynomial time. For this work, van Hoeve et al. received the best paper award at this conference. The results were extended and improved in a subsequent work that is currently under review.

van Hoeve and A. Sabharwal also developed several domain filtering algorithms for constraints involving set variables. In contrast to "traditional variables" that represent a single integer value, set variables represent a set of values. Set variables can be very helpful for modeling purposes, because they often provide a natural way of representing problem entities. Although set variables are present in most constraint programming systems, only few constraints (and corresponding domain filtering algorithms) involving set variables have been implemented. Sabharwal and van Hoeve developed efficient algorithms for two constraints on set variables: the sum-free constraint and the atmost1 constraint. They showed experimentally that both constraints not only offer modeling convenience, but also enable significant computational speed-ups. A paper describing this work has been submitted and is currently under review.

van Hoeve further continued his research on filtering algorithms open constraint programming. A typical application of open constraint programming is task scheduling over alternative resources (e.g., machines). In such case, each resource has its own constraint; for example, tasks assigned to a machine cannot overlap. However, initially it is unknown which task is assigned to which machine, in which case the constraints are said to be "open": the variables on which they are defined are revealed over time. In collaboration with G. Dooms (Brown Univ.), L. Mercier (Brown Univ.), L. Michel (Univ. Connecticut), and P. Van Hentenryck (Brown Univ.), van Hoeve developed new filtering algorithms for several open constraints. The main novelty in their approach was the application of a length-lex ordering to the set variables that represent the scope of an open constraint.

Theme 2 - Mathematical and Computational Foundations of Complex Networks

Over the last year and half, our research under this theme has focused on two main sub-topics:

1. Uncovering structure in patent networks and

2. The study of the impact of network topology on pure Nash equilibrium in graphical games.

Uncovering structure in patent networks

Guo and Gomes are developing of a new methodology to understand innovation through the study of the USPTO Patent database. A central aspect of our approach is the study of the underlying networks induced by the patent and other related data. We can think of the patent data as a network in which patents represent nodes. A given patent A connects to a patent B if patent A cites patent B. The patent data incorporate other networks as well – for example the different entities involved in the process of patenting (e.g., firms, lawyers, patent examiners, etc). Our hypothesis is that we can uncover interesting information about invention and patenting behavior based on the study of the topology and structural properties of the different networks induced by the patent data. In particular, we have developed an approach for automatically ranking patent citations. Our model, SVM Patent Ranking (SVM_{PR}), is based on the large margin SVM classification formulation, incorporating margin constraints that directly capture the specificities of patent ranking. Our approach combines patent-based knowledge features with meta-score features based on several different ad-hoc Information Retrieval methods. Our training algorithm is an anytime algorithm with performance guarantees and is an extension of the Pegasos algorithm, effectively handling hundreds of patents with a few hundred dimensions in the feature space. Experiments on a homogeneous essential wireless patent dataset show that SVM_{PR} performs on average 30%- 40% better than other state-of-the-art ad-hoc Information Retrieval methods, in terms of NDCG measure and a new measure (NMMAP) that we propose in this paper.

A study of the impact of network topology on pure Nash equilibrium in graphical games

In recent years, game theory has moved to the forefront of a number of disciplines: economics, computer science, artificial intelligence, etc. The notion of stability or equilibrium among players is central in game-theoretic settings. A Nash equilibrium is a profile of strategies in which each player has no incentive to deviate from his strategy, given the other players' strategies. Nash proved that every game has a mixed or randomized Nash equilibrium in which a player's strategy is captured by a probability distribution over his actions. On the other hand, when each player has to choose a *pure strategy*, i.e., a single action instead of a randomized mix of actions, a Nash equilibrium is not guaranteed to exist.

Graphical games were proposed as a game-theoretic model for studying large-scale networks of agents with limited interaction. Graphical games naturally occur in markets, the Internet, and numerous settings with non-trivial network topologies. While a mixed Nash equilibrium always exists, deciding whether a pure Nash equilibrium exists is NP-complete for graphical games, even when each player interacts with at most 3 other players. Nevertheless, there are several reasons why pure Nash equilibria are more desirable than mixed Nash equilibria. For example, in economic contexts it has been pointed out that one seldom observes agents resorting to stochastic mechanisms: decision rules used by economic agents may be quite complex but should usually be conceived as deterministic.

A key question concerning PNEs for graphical games is *whether pure Nash equilibria exist, given the topology of the underlying interaction graph*. Known formal results on the existence of PNEs are for random normal form, equivalent to graphical games on complete (clique) graphs. Dilkina, Gomes and Sabharwal study the conditions under which pure Nash equilibria (PNEs) exist for graphical games with random payoffs and show that different interaction graph topologies lead to radically different behavior: depending on the topology, the probability of having a PNE remains quite high or vanishes completely as the number of players grows.

On the theoretical side, Dilkina, Gomes and Sabharwal showed that for games with interaction network consisting of a tree topology the probability of having a PNE converges to zero as the number of players grows. They also demonstrate empirically that the higher the branching factor of the tree, the slower the rate of convergence. Intuitively, while long chains of interactions prevent emergence of PNEs, short paths of interactions between players increases this probability. For instance, for a star topology — one central player interacting with all other players independently —the authors prove that in 2-action games as the number of players grows the probability of having a PNE converges to 0.75. Dilkina, Gomes and Sabharwal further define a family of graphs that generalize the star graph by allowing the center of a graph of *n* nodes to have *m* nodes. The "outer" nodes interact (only) with each of the *m* center nodes, while the center nodes themselves are connected arbitrarily. They refer to this topology as an *m-star*, or equivalently an *augmented bipartite graph*. For such graphs and with *k* possible actions per player, the probability of having a PNE converges to $1-(1-1/k^m)^{k^m}$ as $n \to \infty$ whenever *m* is such that $n/3 - m \to \infty$ (i.e., the center of the graph is small enough). For 2-action games, the lower bound for the probability of having a PNE given this topology lies between $1-1/e \approx 0.63$ and 0.75. Finally, complete bipartite graphs occur naturally in several settings such as when modeling the interactions between buyers and sellers or between bidders and auctioneers. Dilkina, Gomes and Sabharwal show that as the number of players grows, the probability of having a PNE converges to $1-1/e \approx 0.63$. (This is the same value as for standard matrix games, but for a richer underlying topology.)

Empirically, Dilkina, Gomes and Sabharwal studied what happens when a highly regular graph is morphed into a random graph, through a sequence of random re-wirings – a process inspired by the work on Small World Graphs. The experiments showed that the probability of having a PNE increases significantly with more rewiring. A similar increasing probability phenomenon is observed as more and more edges are added to purely random interaction graphs.

In summary, Dilkina, Gomes, and Sabharwal's results demonstrate that the topology of player interactions can greatly affect the probability of existence of a pure Nash equilibrium. Long linear chains of interactions dramatically reduce the chance of having a PNE. One can reverse this effect by adding a small number of random interactions. Such new interactions also implicitly shorten the longest chains in the network. Interestingly, and perhaps contrary to one's intuitions, having more interactions between players can actually increase the probability of having a PNE. Overall, the analysis

suggests that one can exploit the topology of the interaction graph when designing networks of interacting agents or components. Ideally, one would choose a topology that maximizes the probability of having a pure Nash equilibrium, which would be beneficial to all players. The results provide insights for identifying such preferred topologies

Game design

David Schwartz worked with Alex Sarnacki from AFRL/IFSB on a project developing a game prototyping software for studying networked, multiplayer games with intelligent agents. The system would allow "layers" of command hierarchy with human and computer players. By providing this environment, researchers could test algorithms for studying command decisions and visualization on a wide variety of games. The system structure would allow designers to "swap in" different games to study different behaviors.

Starting with one MEng student in Fall 2006 and expanding to three in Spring 2007, David Schwartz worked with a team of students to develop the software using Sarnacki's concepts. He attended the student presentation on 5/11 at Cornell and reported success. David Schwartz continued this work with Alex Sarnacki over the summer as part of his AFRL visiting faculty fellowship work with some assistance of the IISI-supported interns. He also worked on investigating the possibility of having the Asynchronous Chess (AChess) learning framework for adversarial, multi-agent environment being supported on the prototyping framework.

3. Collaborations with AFRL

Asynchronous Chess (AChess) Learning: Learning in a real-time, adversarial, multi-agent environment. Nathaniel Gemelli and Robert Wright (IFSB)

Multi-Agent Sokoban: MAS control and coordination in a computationally complex logistics domain. James Lawton (IFSB)

Automated Reasoning: n-Queens Completion Problem Andrew Boes (IFSB)

Efficient Mission-based Information Retrieval Pete Lamonica. (IFED)

FLEXDB: An Efficient, Scalable and Secure Peer-to-Peer XML Database. Jim Nagy. (IFED)

Information Extraction; Mark Zappavigna, Jeff Hudack (IFED)

Knowledge-based inference. Mark Zappavigna, Jeff Hudack. (IFED)

Wargame design, David Ross (IFSB)

SimBionic for wargame development. David Ross (IFSB)

WARCON (working title) software for Air Academy David Ross, IFSD

4. Personnel

IISI Members:

Claire Cardie (Computer Science)
Rich Caruana (Computer Science)
Jon Conrad (Applied Economics and Management)
Raffaello D'Andrea (Mechanical & Aerospace Engineering)
Johannes Gehrke (Computer Science)
Carla Gomes (Computer Science)
Joseph Halpern (Computer Science)
Juris Hartmanis (Computer Science)
John Hopcroft (Computer Science)
Thorsten Joachims (Computer Science)
Jon Kleinberg (Computer Science)
Lillian Lee (Computer Science)
William Lesser (Applied Economics and Management)
Jose F. Martinez (Electrical & Computer Engineering)
Venkatesh Rao (Mechanical & Aerospace Engineering)
Ashish Sabharwal (Computer Science)
David Schwartz (Computer Science)
Bart Selman (Computer Science)
Phoebe Sengers (Information Science)
David Shmoys (Operations Research)
Chris Shoemaker (Civil Engineering)
Steve Strogatz (Theoretical and Applied Mechanics)
Willem-Jan van Hoeve (Computer Science)
Stephen Wicker (Electrical and Computer Engineering)

IISI Visitors:

Michele Lombardi (University of Bologna)

Short visits:
Alon Altman (Technion)

Eyal Amir (University of Illinois at Urbana-Champaign)
Michael J. Black (Brown University)
David Blei (Princeton University)
Sandra Carberry (University of Delaware)
Aron Culotta (University of Massachusetts at Amherst)
Yolanda Gil (University of Southern California)
Joerg Hoffmann (University of Innsbruck, Austria)
Rebecca Hwa (University of Pittsburgh)
Bo Pang (Yahoo! Research)
Stefan Pohl (Universitaet Darmstadt)
Jonathan Victor (Weill Medical College)
Xiaoyan Zhu (Tsinghua University)
Larry Zitnick (Microsoft Research)

5. Publications

Publications by IISI Members:

Basu, A., N. Kırman, M. Kırman, M. Chaudhuri, and J.F. Martínez. Scavenger: A new last level cache architecture with global block priority. In Intl. Symp. on icroarchitecture, Chicago, IL, Dec. 2007.

Breck, Eric, Yejin Choi, and Claire Cardie. Identifying Expressions of Opinion in Context. Twentieth International Joint Conference on Artificial Intelligence (IJCAI), 2007.

Breck, Eric, Yejin Choi, Veselin Stoyanov, and Claire Cardie. Cornell System Description for the NTCIR-6 Opinion Task. The 6th NTCIR Workshop Meeting, Tokyo, Japan, 2007.

Cardie, Claire, Cynthia Farina, Thomas Bruce, and Erica Wagner. Using Natural Language Processing to Improve E-rulemaking. Proceedings of the 7th Annual International Conference on Digital Government Research, 2006.

Cardie, Claire, Cynthia Farina, Matt Rawding, Adil Aijaz. An eRulemaking Corpus: Identifying Substantive Issues in Public Comments. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 2008. (to appear)

Cardie, Claire, Cynthia Farina, Adil Aijaz, Matt Rawding, Stephen Purpura. A Study in Rule-Specific Issue Categorization for e-Rulemaking. 9th Annual International Conference on Digital Government Research, Montreal, Canada, 2008. (to appear)

Choi, Yejin and Claire Cardie. Structured Local Training and Biased Potential Functions for Conditional Random Fields with Application to Coreference Resolution. NAACL Human Language Technology Conference (NAACL-HLT), 2007.

Choi, Yejin, Eric Breck, and Claire Cardie. Joint Extraction of Entities and Relations for Opinion Recognition. Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2006.

Conrad, Jon, Carla P. Gomes, Willem-Jan van Hoeve, Ashish Sabharwal, Jordan Suter. Connections in Networks: Hardness of Feasibility versus Optimality. CPAIOR-07. 4th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, pp 16-28, Brussels, Belgium, May 2007.

Dilkina, Bistra, Carla P. Gomes, Ashish Sabharwal. The Impact of Network Topology on Pure Nash Equilibria in Graphical Games. AAAI-07. 22nd Conference on Artificial Intelligence, pp 42-49, Vancouver, BC, Canada, July 2007. Nominated for the Best Paper Award. Also in NESCAI-07 (preliminary version)

Dilkina, Bistra, Carla Gomes, and Ashish Sabharwal. Tradeoffs in the Complexity of Backdoor Detection. Proc. 13th International Conference on Principles and Practice of Constraint Programming, Providence, RI, Sep 2007.

Farina, Cynthia, Claire Cardie, Thomas Bruce, Erica Wagner. Better Inputs for Better Outcomes: Using the Interface to Improve e-Rulemaking. Workshop on eRulemaking at the Crossroads, Proceedings of the 7th Annual International Conference on Digital Government Research, 2006.

Gomes, Carla, Willem van Hoeve, Ashish Sabharwal, and Bart Selman. Counting CSP Solutions Using Generalized XOR Constraints. AAAI-07. Proceedings of the 22nd National Conference on Artificial Intelligence, Vancouver, Canada, Jul 2007.

Gomes, Carla, Ashish Sabharwal, and Bart Selman. From Sampling to Model Counting. IJCAI07. Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, Jan 2007. (Nominated for Best Paper Award).

Gomes, Carla, Ashish Sabharwal, and Bart Selman. Near-Uniform Sampling of Combinatorial Spaces Using XOR Constraints. NIPS-06. Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, B. C., Dec 2006

Gomes, Carla P., Joerg Hoffmann, Ashish Sabharwal, Bart Selman. Short XORs for Model Counting: From Theory to Practice. SAT-07. 10th International Conference on

Theory and Applications of Satisfiability Testing, pp 100-106, Lisbon, Portugal, May 2007. (short paper).

Gomes, Carla P., Joerg Hoffmann, Ashish Sabharwal, Bart Selman. From Sampling to Model Counting. IJCAI-07. 20th International Joint Conference on Artificial Intelligence, pp 2293-2299, Hyderabad, India, Jan 2007. Nominated for the Best Paper Award.

Gomes, C.P., W.-J. van Hoeve, and A. Sabharwal. Connections in Networks: A Hybrid Approach. In Proceedings of the Fifth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR 2008) (short paper), 2008.

Hopcroft, John, Anirban Dasgupta and Ravi Kannan. Spectral Clustering by Recursive Partitioning. ESA 2006.

Hopcroft, John, Anirban Dasgupta and Ravi Kannan. Spectral Clustering with Limited Independence. SODA 2007.

Kroc, Lukas, Ashish Sabharwal, Bart Selman. Survey Propagation Revisited. UAI-07. 23rd Conference on Uncertainty in Artificial Intelligence, pp 217-226, Vancouver, BC, Canada, July 2007. Nominated for the Best Student Paper Award.

Kroc, Lukas, Ashish Sabharwal, Bart Selman. Leveraging Belief Propagation, Backtrack Search, and Statistics for Model Counting. CPAIOR-08: 5th International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, pp 127-141, Paris, France, May 2008.

Kroc, Lukas, Ashish Sabharwal, Bart Selman. Comparing Message-Passing and Local Heuristics in Decimation Strategies. Under review, 2008.

Kroc, Lukas, Ashish Sabharwal, Bart Selman. Counting Solution Clusters Using Belief Propagation. Under review, 2008.

Stoyanov, Veselin and Claire Cardie. Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning. Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2006.

Stoyanov, Veselin and Claire Cardie. Toward Opinion Summarization: Linking the Sources. COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text, 2006.

Stoyanov, Veselin and Claire Cardie. Annotating Topics of Opinions. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 2008. (to appear)

van Es, Harold, Carla Gomes, Meinolf Sellmann, and Cindy van Es. Spatially-Balanced Complete Block Designs for Field Experiments. Geoderma Journal, Vol. 140 (2007) 346--352.

van Hoeve, Willem, Carla Gomes, Michele Lombardi, and Bart Selman. Optimal Multi-Agent Scheduling with Constraint Programming. In Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI 2007), 2007.

van Hoeve, Willem-Jan, Gilles Pesant, Louis-Martin Rousseau, Ashish Sabharwal. Revisiting the Sequence Constraint. CP-06. 12th International Conference on Principles and Practice of Constraint Programming, LNCS volume 4204, pp 620-634, Nantes, France, Sep 2006. Best Paper Award.

van Hoeve, Willem-Jan, Gilles Pesant, Louis-Martin Rousseau, Ashish Sabharwal. New Filtering Algorithms for Combinations of Among Constraints. Under review, 2007.

van Hoeve, W.-J. and A. Sabharwal. Filtering Atmost1 on Pairs of Set Variables. In Proceedings of the Fifth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR 2008) (short paper), 2008.

van Hoeve, W.-J., C.P. Gomes, M. Lombardi, and B. Selman. Optimal Multi-Agent Scheduling with Constraint Programming. In Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI 2007), 2007.

van Hoeve, W.-J. and A. Sabharwal. Two Set-Constraints for Modeling and Efficiency. In Proceedings of the 6th International Workshop on Constraint Modelling and Reformulation (ModRef 2007), 2007.

6. Meetings

IISI organized and sponsored the following events in 2007-08:

NESCAI 07, http://www.cs.cornell.edu/Conferences/nescai/nescai07/
NESCAI 08, http://www.cs.cornell.edu/Conferences/Nescai/

IISI participated in the following events in 2007-08:

Conferences/Workshops:

- Twenty-Second National Conference on Artificial Intelligence (AAAI-07)
- Twenty-Third National Conference on Artificial Intelligence (AAAI-08)
- Thirteenth International Conference on Principles and Practice of Constraint Programming (CP 2007)
- Fourth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR 2007)

- Fifth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR 2008)
- Tenth International Symposium on Artificial Intelligence and Math (ISAIM 2008)
- Bits On Our Minds (BOOM 2008)
- Cornell University Artificial Intelligence Seminar (2007, 2008)